# Study on Page Ranking Algorithms for Search Engine Optimization

Tin Tin Yu
baybayyu@gmail.com
University of Computer Studies, Mandalay

## Abstract

***As the enormous increase of the World Wide Web, search engine become important and play an essential part of the retrieving information. Search Engine provides the gateway for most of the users trying to explore the huge information base of web pages. The purpose of web search engines is to return web pages lists that are relevance to the user query. The problem with web search relevance ranking is to estimate the resulted list is how much relevance of a page with the user query. Since Search Engine Optimization (SEO) has become an important part of search engine marketing, the evolution of page rank algorithms have been focused in information retrieval research field. Page Rank is a topic much discussed by SEO experts. The most popular page ranking algorithms have been reviewed in this paper.***

***Keyword: World Wide Web, Search Engine, Search Engine Optimization, Page ranking Algorithms.***

## 1. Introduction

World Wide Web provides us with heavy amount of necessary data digitally available as hypertext Data may be WebPages, images, information and other types. This hypertext pool is dynamically changing due to this reason it is more difficult to find useful information. The economic importance of web will enhance the academic interest. The database administrator, management persons or others wishing to perform data mining on large number of web pages will require the services of web data mining tools and techniques.

Clustering techniques become one of the alternative solutions for the problem of search engine [5]. Clustering provided an organized way to manage our search engine. Cosine similarity is also widely used in text or documents clustering. Ranking of the documents is done using their similarity values. The top ranked documents are regarded as more relevant to the query.

The issues of improving search engines have been solved by employing classification. It is managed by the following issues: 1) Information growth on the web is increasing at an exponential rate, 2) discovering structure within an information domain has been proved to be an effective way in practice of organizing large data sets. Classification deals with such type of problem that the retrieved results from traditional search engines are topic-independent. Most previous work had used class size or the alphabet of the class label to rank the class[7].

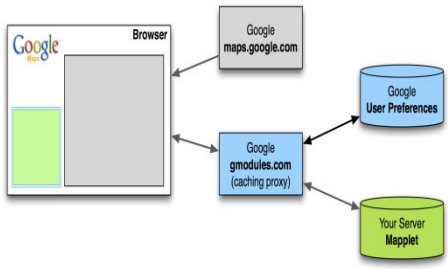


**Figure 1: High level Google Architecture**

## 2. Related Work

There has been considerable research on ranking results. We collect related work in this section to provide an overview of web search engine and ranking algorithms that are the most relevant or most closely related the content presented in this paper.

Search research on the web has a short and concise history. The World Wide Web Worm (WWWW)[7] was one of the first web search engines. It was subsequently followed by several other academic search engines, many of which are now public companies. Compared to the growth of the Web and the importance of search engines there are precious few documents about recent search engines [6]. According to Michael Mauldin (chief scientist, Lycos Inc) [3], "the various services (including Lycos) closely guard the details of these databases". However, there has been a fair amount of work on specific features of search engines. Especially well represented is work which can get results by post-processing the results of existing commercial search engines, or produce small scale "individualized" search engines. Finally, there has been a lot of research on information retrieval systems, especially on well controlled collections.

Information Retrieval (IR) Systems are the predecessors of Web and search engines. These systems were designed to retrieve documents in curate digital collections such as library abstracts, corporate documents, news, etc. Traditionally, IR relevance ranking algorithms were designed to obtain high recall on medium-sized document collections using long detailed queries. Furthermore, textual documents in these collections had little or no structure or hyperlinks. Web search engines incorporated many of the principles and algorithms of Information Retrieval Systems, but had to adapt and extend them to fit their needs. Early Web Search engines such as Lycos and AltaVista concentrated on the scalability issues of running web search engines using traditional relevance ranking algorithms. Newer search engines, such as Google, exploited web-specific relevance features such as hyperlinks to obtain significant gains in quality. These measures were partly motivated by research in citation analysis carried out in the bibliometrics field [9].

PageRank is one of the methods Google uses to determine a page's relevance or importance. It is only one part of the story when it comes to the Google listing[10].

## 3. Web Search Engine

We use Search Engines to search for information across the Internet. Internet being an ever-expanding ocean of data, their importance grew with every passing day. The diversity of the information itself made it necessary to have a tool to cut down on the time spent in searching.According to functioning, there are three types of search engine [2].

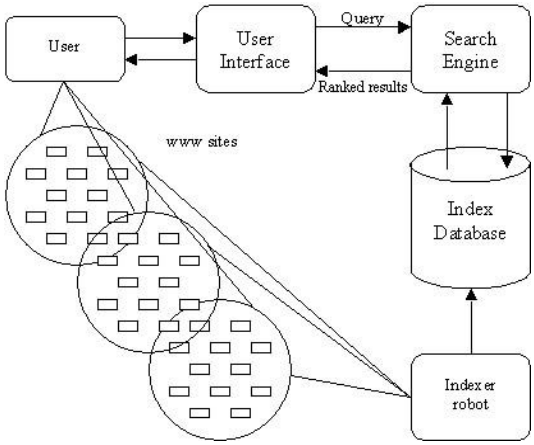


**Figure 2: Architecture of Search Engine**

When user types something in the search engine box the search engine processes the user request by matching the user query with the results stored in the database. The results are stored in the database in the form of web pages. These web pages are ranked on the basis of content of the web page, relevant keywords used in the web pages, the frequency of occurring of keywords in the web page . If the title, description, content of the web page is more relevant and important then web pages are listed at the top. The title or description of the web sites should appear to user as the useful link because the users normally visit or attempt to click on the web pages that are given at the top. The web pages are ranked on basis of the numbers assigned to these web pages. The web pages are stored in the database and retrieved with help of result engine. Results are displayed on the user screen in the form of a list from the most relevant to least relevant web sites [3]. These are the technical aspects of SEO. Page ranking algorithms are used and reivse to produce the more optimized accurate and relevant resulted list of user query. In the rest of paper, we discuss the popular page ranking alogrithm.

## 4. Page Ranking Algorithms

The World Wide Web contains an enormous amount of information, but it can be exceedingly difficult for users to locate resources that are both high in quality and relevant to their information needs. Issues that have to be dealt with are the detection of relevant information, involving the searching and indexing of the Web content, the creation of some meta knowledge out of the information which is available on the Web, as well as the addressing of the individual users' needs and interests, by personalizing the provided information and services. In this paper we discuss mainly two algorithms and other related ranking algorithms [1].

Search engines use two different kinds of ranking factors: query dependent factors and query Independent Factors .Query-dependent are all ranking factors that are specific to a given query, while query-independent factors are attached to the documents, regardless of a given query. Query-dependent factors used by search engines are measures such as word documents frequency, the position of the query terms within the document or the inverted document frequency, which are all measures that are used in traditional Information Retrieval. Some of the query independent factors are Link popularity, Click popularity and up to dateness etc. Ranking algorithms based on link popularity, falls under Link based ranking algorithm category[8].

Two popular webpage ranking algorithms are HITS and PageRank. HITS emphasizes mutual reinforcement between authority and hub web pages, while PageRank emphasizes hyperlink weight normalization and web surfing based on random walk models.

### 4.1. Hypertext Induced Topics Search

Hypertext Induced Topics Search (HITS) is developed by Jon Kleinberg. HITS is applied on a sub graph after a search is done on the complete graph. It uses hubs and authorities to define a recursive relationship between web pages. The algorithm performs a series of iterations, each consisting of two basic steps:

**Step1**- Authority Update: Update each node's Authority score to be equal to the sum of the Hub Scores of each node that points to it. That is, a node is given a high authority score by being linked to by pages that are recognized as Hubs for information.

**Step2**- Hub Update: Update each node's Hub Score to be equal to the sum of the Authority Scores of each node that it points to. That is, a node is given a high hub score by linking to nodes that are considered to be authorities on the subject.

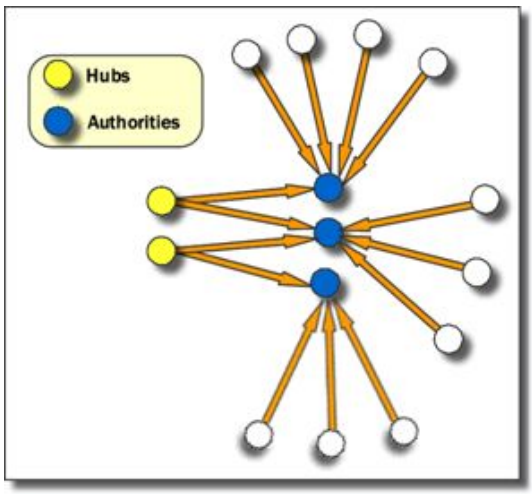


**Figure 3: Hubs and Authorities**

HITS has the ability of ranking page according to the query topic. Because of this fact, HITS can give the result list that are more relevant with the user query, in other words, more relevant authority and hub pages.This type of ranking may also be combined with the information retrieval based ranking techniques.

And also HITS algorithm has some limitation. First of all, it does not have the anti-spam capability of PageRank. It is quite easy to influence HITS by adding out-links from one's own page to point to many good authorities. This boosts the hub score of the page. Because hub and authority scores are interdependent, it in turn also increases the authority score of the page. Another problem of HITS is topic drift. In expanding the root set, it can easily collect many pages (including authority pages and hub pages) which have nothing to do the search topic because out-links of a page may not point to pages that are relevant to the topic and in-links to pages in the root set may be irrelevant as well because people put hyperlinks for all kinds of reasons, including spamming. The query time evaluation is also a major drawback. Getting the root set, expanding it and then performing eigenvector computation are all time consuming operation[2].

### 4.2. PageRank Algorithm

PageRank is an objective measure of its citation importance that corresponds well with people's subjective idea of importance. Because of this correspondence, PageRank is an excellent way to prioritize the results of web keyword searches. For most popular subjects, a simple text matching search that is restricted to web page titles performs admirably when PageRank prioritizes the results. For the type of full text searches in the main Google system, PageRank also helps a great deal.

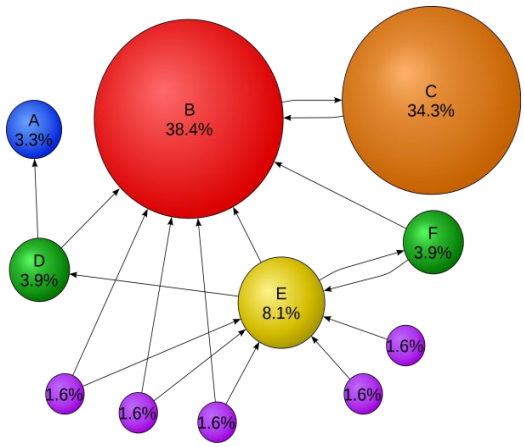


**Figure 4: Simplified diagram illustrating a simple search engine utilizing the standard PageRank**

The main advantage of PageRank is its ability to fight spam. Recognizing and fighting spam is an important issue in Web search. A page is important if the pages pointing to it are important. Since it is not easy for Web page owner to add in-links into his/her page from other important pages, it is thus not easy to influence PageRank.As pageRank is a query independent algorithm i.e. it precomputes the rank score.So it takes very less time .Both these two advantages contributed greatly to Google's success. And also It returns important pages as Rank is calculated on the basis of the popularity of a page. For calculating rank value of a page, it consider the entire web graph, rather than a small subset, it is less susceptible to localized link spam.

The main disadvantage is that it favors older pages, because a new page, even a very good one, will not have many links unless it is part of an existing site. PageRank the web pages are ranked according to the number of clicks made on that particular web page but this may lead to illegal ranking of web pages . In other word, whenever a query is given the pages that are satisfying the query are presented according to the rank of the page. The top most one will be given highest priority. The highest priority is because the number of clicks on that particular web page are more without concerned with the content that is present in that particular web page. For this purpose the ranking should be given according to the content present in the web page rather than the number of clicks made on that particular web page. Because if a wrong page is presented to end user then he will browse the page which will increase the click count of the traced page which is wrong. This leaves the web page with highest priority. For this purpose it is better to rank pages according to the content in the web page. This leads to the combination of text mining with web mining.

## 5.Conclusion

Each ranking algorithm provides a definite rank to resultant web pages. The difference betweeen rankings producted by difference alogrithms reflects slightly different but useful emphasis. Basically, indegree and outdegree are fundamental important in web ranking. To get the more optimized result for the user, a typical search engine should use web page ranking techniques and algorithm according to the specific needs of the users. Since search engine is a complex system , the further enhancements should be focus continuously. After the study analysis of ranking algorithms, we can conclude that an efficient web page ranking algorithm should be able to solve the challenges and limitation of existing application efficiently and it should compact with global standards of web technology in terms of accuracy, relevancy and response time of the results.

## References


[1]Bing Lu, “Web Data Mining”, Exploring Hyperlinks, Contents, and Usage Data, Department of Computer Science, University of Illinois, Chigao, liub@cs.uic.edu

[2]I.va Gregurec, Petna Grd “Search engine optimization website analysis of selected faculties in Croatia”. Central European conference on information and intelligent systems, sep 19-21, 2012 PP-213.

[3]Mauldin, Michael L. Lycos “Design Choices in an Internet Search Service”, IEEE Expert Interview http://www.computer.org/pubs/expert/1997/trends/x1008/mauldin.htm

[4]McBryan 94, Oliver A. McBryan. GENVL and WWWW: “Tools for Taming the Web. First International Conference on the World Wide Web”,CERN, Geneva (Switzerland), May 25-26-27, 1994.

[5]Minky Jindal and Nisha Kharb ,”k-mean s clustering technique on search engine dataset using data mining tool”, CSE/IT Department, ITM University, Sector-23A, Gurgaon, India.

http://www.cs.colorado.edu/home/mcbryan/ypapers/www94.ps

[6]Pinkerton 94, Brian Pinkerton, “Finding What People Want: Experiences with the WebCrawler”., The Second International WWW Conference Chicago, USA, October 17-20, 1994.

[7]Rajesh Singh and Ajmer (Rajasthan), “An approach for search engine optimizationusing classification - a data mining technique ”, Scholar inBhagvant University, S.K. Gupt, A.P.,CSED, BIET, Jhansi (U.P.)

[8]http://searchengineland.com/guide/seo/types- of-search-engine-ranking-factors

[9]http://www.wordstream.com/articles/internet - search-engines-history

[10]http://www.db.stanford.edu/%7Ebackrub/google.html